(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－182685
（P2002－182685A）
(43)公開日　平成14年６月26日(2002.6.26)

(51)Int.Cl.⁷	識別記号	ＦＩ	テーマコード（参考）
Ｇ１０Ｌ　15/18		Ｇ０６Ｔ　7/00	Ｐ　５Ｄ０１５
Ｇ０６Ｔ　7/00			３００Ｆ　５Ｌ０９６
	３００	Ｇ１０Ｌ　3/00	５３７Ｚ
Ｇ１０Ｌ　15/00			５５１Ｈ
15/24			５７１Ｑ

審査請求　未請求　請求項の数12　ＯＬ　（全 10 頁）

(21)出願番号　特願2000－376911(P2000－376911)

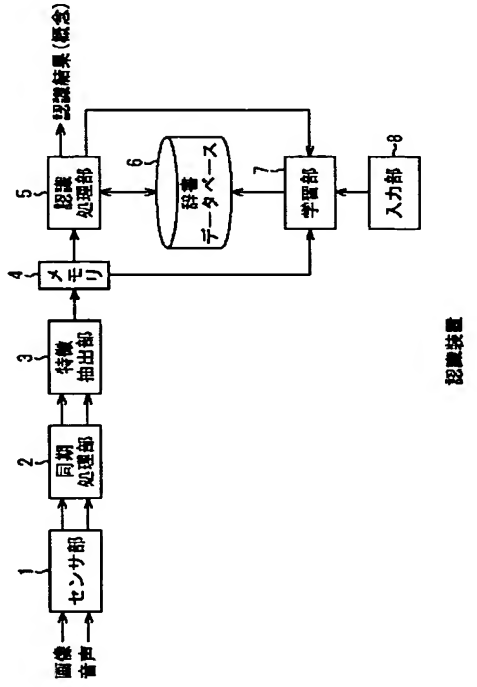
(22)出願日　平成12年12月12日(2000.12.12)

(71)出願人　000002185
ソニー株式会社
東京都品川区北品川６丁目７番35号
(72)発明者　中塚　洪長
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(74)代理人　100082131
弁理士　稲本　義雄

Ｆターム(参考)　5D015 FF00 HH04
5L096 BA05 FA23 GA41 HA04 JA11 KA04

(54)【発明の名称】　認識装置および認識方法、学習装置および学習方法、並びに記録媒体

(57)【要約】
【課題】　認識性能を向上させる。
【解決手段】　同期処理部２は、入力された画像と音声を同期させ、特徴抽出部３は、その同期された画像と音声それぞれから、特徴量を抽出して、その画像と音声の特徴量を合成した合成特徴量を得る。学習部７は、その合成特徴量に基づいて学習を行い、同一概念を表す画像および音声に対応するモデルを生成し、そのモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する。一方、認識処理部５は、合成特徴量と、辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する。



1

【特許請求の範囲】

【請求項１】　同一概念を表す画像および音声のモデルと、その概念を表す概念情報とを対応付けた辞書を記憶する記憶手段と、

入力された画像と音声を同期させる同期手段と、

同期された前記画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出手段と、

前記抽出手段において出力される合成特徴量と、前記辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する認識手段とを備えることを特徴とする認識装置。

【請求項２】　前記同期手段は、

入力された画像の区間である画像区間と、入力された音声の区間である音声区間を検出し、

前記画像区間と音声区間それぞれを正規化することにより、入力された画像と音声を同期させることを特徴とする請求項１に記載の認識装置。

【請求項３】　前記同期手段は、さらに、正規化された前記画像区間と音声区間それぞれの始点と終点とを一致させることにより、入力された画像と音声を同期させることを特徴とする請求項２に記載の認識装置。

【請求項４】　前記同期手段は、さらに、正規化された前記画像区間と音声区間それぞれにおける画像のフレームと音声のフレームとを対応させることにより、入力された画像と音声を同期させることを特徴とする請求項２に記載の認識装置。

【請求項５】　同一概念を表す画像および音声のモデルと、その概念を表す概念情報とを対応付けた辞書を参照して認識処理を行う認識方法において、

入力された画像と音声を同期させる同期ステップと、

同期された前記画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、

前記抽出ステップにおいて出力される合成特徴量と、前記辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する認識ステップとを備えることを特徴とする認識方法。

【請求項６】　同一概念を表す画像および音声のモデルと、その概念を表す概念情報とを対応付けた辞書を参照して行う認識処理を、コンピュータに行わせるプログラムが記録されている記録媒体において、

入力された画像と音声を同期させる同期ステップと、

同期された前記画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、

前記抽出ステップにおいて出力される合成特徴量と、前記辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する認識ステップとを備えるプログラムが記録されていることを特徴とする記録媒体。

2

【請求項７】　入力された画像と音声を同期させる同期手段と、

同期された前記画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出手段と、

前記抽出手段において出力される合成特徴量に基づいて学習を行うことによりモデルを生成し、同一概念を表す画像および音声に対応するモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する学習手段とを備えることを特徴とする学習装置。

【請求項８】　前記同期手段は、

入力された画像の区間である画像区間と、入力された音声の区間である音声区間を検出し、

前記画像区間と音声区間それぞれを正規化することにより、入力された画像と音声を同期させることを特徴とする請求項７に記載の学習装置。

【請求項９】　前記同期手段は、さらに、正規化された前記画像区間と音声区間それぞれの始点と終点とを一致させることにより、入力された画像と音声を同期させることを特徴とする請求項８に記載の学習装置。

【請求項１０】　前記同期手段は、さらに、正規化された前記画像区間と音声区間それぞれにおける画像のフレームと音声のフレームとを対応させることにより、入力された画像と音声を同期させることを特徴とする請求項８に記載の学習装置。

【請求項１１】　入力された画像と音声を同期させる同期ステップと、

同期された前記画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、

前記抽出ステップにおいて出力される合成特徴量に基づいて学習を行うことによりモデルを生成し、同一概念を表す画像および音声に対応するモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する学習ステップとを備えることを特徴とする学習方法。

【請求項１２】　所定の学習処理を、コンピュータに行わせるプログラムが記録されている記録媒体において、

入力された画像と音声を同期させる同期ステップと、

同期された前記画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、

前記抽出ステップにおいて出力される合成特徴量に基づいて学習を行うことによりモデルを生成し、同一概念を表す画像および音声に対応するモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する学習ステップとを備えるプログラムが記録されていることを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、認識装置および認識方法、学習装置および学習方法、並びに記録媒体に関し、特に、例えば、画像と音声から、その概念を認識する場合において、高い認識性能を得ることができるようにする認識装置および認識方法、学習装置および学習方法、並びに記録媒体に関する。

【０００２】

【従来の技術】近年においては、CPUの高速化、メモリの大容量化等が進み、例えば、音声認識装置を搭載したエンタテイメント用のロボット等が低価格で実現されている。

【０００３】このようなロボットは、ユーザの音声を音声認識し、その認識結果に基づいて、各種の行動を起こす。即ち、例えば、ユーザが、「お手」と発話した場合には、ロボットは、実際の犬がお手をするような行動を起こす。

【０００４】

【発明が解決しようとする課題】ところで、例えば、ロボットにおいて、外部から与えられる刺激としての音声だけでなく、画像についても処理を行い、その音声と画像が表す概念を認識して、その認識した概念に基づいて、行動するようにすれば、よりエンタテイメント性の向上させたロボットを実現することができると予想される。

【０００５】しかしながら、例えば、ロボットに対して、ある概念を表す音声と画像が与えられる場合、その音声と画像とは、同期したものとはなっていないため、そのような同期していない画像と音声を処理して、その画像と音声が表す概念を認識しても、十分な認識性能が得られないことが予想される。

【０００６】また、上述のように、画像と音声が表す概念を認識するには、あらかじめ学習を行っておく必要があるが、学習に際しても、ある概念を表す音声と画像が、同期したものとなっていない場合には、十分な認識性能が得られないことが予想される。

【０００７】本発明は、このような状況に鑑みてなされたものであり、画像と音声から、その概念を認識する場合において、高い認識性能を得ることができるようにするものである。

【０００８】

【課題を解決するための手段】本発明の認識装置は、入力された画像と音声を同期させる同期手段と、同期された画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出手段と、抽出手段において出力される合成特徴量と、辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する認識手段とを備えることを特徴とする。

【０００９】本発明の認識方法は、入力された画像と音声を同期させる同期ステップと、同期された画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、抽出ステップにおいて出力される合成特徴量と、辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する認識ステップとを備えることを特徴とする。

【００１０】本発明の第１の記録媒体は、入力された画像と音声を同期させる同期ステップと、同期された画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、抽出ステップにおいて出力される合成特徴量と、辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する認識ステップとを備えるプログラムが記録されていることを特徴とする。

【００１１】本発明の学習装置は、入力された画像と音声を同期させる同期手段と、同期された画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出手段と、抽出手段において出力される合成特徴量と、合成特徴量に基づいて学習を行い、同一概念を表す画像および音声から得られる合成特徴量と、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する学習手段とを備えることを特徴とする。

【００１２】本発明の学習方法は、入力された画像と音声を同期させる同期ステップと、同期された画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、抽出ステップにおいて出力される合成特徴量に基づいて学習を行うことによりモデルを生成し、同一概念を表す画像および音声に対応するモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する学習ステップとを備えることを特徴とする。

【００１３】本発明の第２の記録媒体は、入力された画像と音声を同期させる同期ステップと、同期された画像と音声それぞれから、特徴量を抽出し、その画像と音声の特徴量を合成した合成特徴量を出力する抽出ステップと、抽出ステップにおいて出力される合成特徴量に基づいて学習を行うことによりモデルを生成し、同一概念を表す画像および音声に対応するモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する学習ステップとを備えるプログラムが記録されていることを特徴とする。

【００１４】本発明の認識装置および認識方法、並びに第１の記録媒体においては、入力された画像と音声が同期され、その同期された画像と音声それぞれから、特徴量が抽出され、その画像と音声の特徴量を合成した合成特徴量が出力される。そして、その合成特徴量と、辞書におけるモデルとを用いてマッチングを行うことによ

り、入力された画像と音声が表す概念が認識される。

【００１５】本発明の学習装置および学習方法、並びに第２の記録媒体においては、入力された画像と音声が同期され、その同期された画像と音声それぞれから、特徴量が抽出され、その画像と音声の特徴量を合成した合成特徴量が出力される。そして、その合成特徴量に基づいて学習が行われることによりモデルが生成され、同一概念を表す画像および音声に対応するモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書が生成される。

【００１６】

【発明の実施の形態】図１は、本発明を適用した認識装置の一実施の形態の構成例を示している。

【００１７】センサ部１は、例えば、少なくとも、音声（音）を集音して出力するマイク（マイクロフォン）と、画像を撮影して出力するビデオカメラから構成され、外部から与えられる刺激としての音声と画像を感知し、Ａ／Ｄ(Analog Digital)変換して、対応する音声データと画像データを、同期処理部２に出力する。

【００１８】同期処理部２は、センサ部１が出力する音声データと画像データについて、後述するような同期処理を施し、これにより、相互に同期した音声データと画像データを、所定のフレーム単位で、特徴抽出部３に供給する。

【００１９】特徴抽出部３は、同期処理部２から供給される音声データを、そのフレーム（以下、適宜、音声フレームという）単位で処理し、音声の特徴パラメータを抽出する。さらに、特徴抽出部３は、同期処理部２から供給される画像データを、そのフレーム（以下、適宜、画像フレームという）単位で処理し、画像の特徴パラメータを抽出する。そして、特徴抽出部３は、音声フレームから得られた音声の特徴パラメータと、その音声フレームに対応する画像フレームから得られた画像の特徴パラメータとを合成し、その結果得られる合成特徴パラメータを、メモリ４に供給する。

【００２０】なお、音声と画像の特徴パラメータの合成方法としては、例えば、その音声と画像の特徴パラメータがベクトルで構成される場合に、その音声と画像の特徴ベクトルの要素（コンポーネント）を接続して１つのベクトルを構成する方法等がある。

【００２１】また、音声の特徴パラメータとしては、例えば、ＭＦＣＣ(Mel Frequency Cepstrum Coefficients)や、そのフレーム間差分、パワー等を用いることができる。さらに、画像の特徴パラメータとしては、動きベクトルや、色情報、ＤＣＴ(Discrete Cosine Transform)係数、画像に表示された物体の形状を表す情報等を用いることができる。

【００２２】メモリ４は、特徴抽出部３から供給される合成特徴パラメータを一時記憶する。

【００２３】認識処理部５は、辞書データベース６の辞書に登録されている各モデルと、メモリ４に記憶された合成特徴パラメータとを用いて、各モデルから、合成特徴パラメータが観測されるスコア（尤度）を求めるマッチング処理を行い、その合成特徴パラメータと最もマッチするモデル、即ち、例えば、最高スコアのモデルを求める。さらに、認識処理部５は、その最高スコアに基づいて、いま認識の対象となっている音声および画像に対応するモデルが、既に、辞書データベース６の辞書に登録されているかどうか、即ち、学習済みかどうかを判定し、学習済みの場合には、最高スコアのモデルに対応する概念情報を求め、その概念情報を、センサ部１に入力された音声と画像が表す概念の認識結果として出力する。また、認識処理部５は、いま認識の対象となっている音声および画像に対応するモデルが学習済みでない場合（いま認識の対象となっている音声および画像に対応するモデルが、辞書データベース６の辞書に登録されていない場合）、学習部７に対して、そのモデルの学習を要求する。

【００２４】なお、ここでは、スコアは、その値が大きいほど尤度が高くなるものとする。また、スコアとしては、例えば、モデルから合成特徴パラメータが観測される確率や、合成特徴パラメータのモデルに対する類似性（例えば、特徴空間におけるモデルと合成特徴パラメータとの距離）(Confidence Measure)等を用いることが可能である。但し、スコアとして、例えば、特徴空間におけるモデルと合成特徴パラメータとの距離を用いる場合には、スコアの値が小さいほど、尤度が高いことを意味することになる。

【００２５】辞書データベース６は、合成特徴パラメータを用いて学習を行うことにより得られる音声および画像のモデルと、その音声および画像の概念を表す概念情報とが対応付けられて登録される辞書を記憶している。なお、音声および画像のモデルとしては、例えば、ＨＭＭ(Hidden Markov Model)等の確率モデルその他を採用することができる。

【００２６】学習部７は、認識処理部５から学習の要求があると、メモリ４に記憶された合成特徴パラメータを用いて、音声および画像のモデルの学習を行う。さらに、学習部７は、学習の結果得られたモデルを、そのモデルが表す音声および画像の概念情報と対応付けて、辞書データベース６の辞書に登録する。

【００２７】入力部８は、例えば、キーボード等で構成され、学習部７が学習の対象としているモデルに対応する音声および画像の概念情報を入力するときに、ユーザによって操作される。入力部８が操作されることにより入力された概念情報は、学習部７に供給される。なお、入力部８は、その他、例えば、音声認識装置等で構成することも可能であり、この場合、概念情報は、音声で入力することができる。

【００２８】次に、図２のフローチャートを参照して、

図１の認識装置の動作について説明する。

【００２９】センサ部１は、音声と画像を感知し、対応する音声データと画像データを、同期処理部２に出力する。同期処理部２は、ステップＳ１において、センサ部１が出力する音声データと画像データについて、同期処理を施し、これにより、相互に同期した、所定の音声フレームごとの音声データと、所定の画像フレームごとの画像データを、特徴抽出部３に供給する。

【００３０】特徴抽出部３は、ステップＳ２において、同期処理部２からの音声フレーム単位の音声データから、その特徴パラメータを抽出するとともに、同じく、同期処理部２からの画像フレーム単位の画像データから、その特徴パラメータを抽出する。さらに、特徴抽出部３は、ステップＳ２において、各音声フレームから得られた音声の特徴パラメータと、その音声フレームに対応する画像フレームから得られた画像の特徴パラメータとを合成し、合成特徴パラメータとして、メモリ４に供給して記憶させる。

【００３１】そして、ステップＳ３に進み、認識処理部５は、マッチング処理を行い、スコアを求める。即ち、認識処理部５は、辞書データベース６の辞書に登録されている各モデルと、メモリ４に記憶された合成特徴パラメータとを用いて、各モデルから、合成特徴パラメータが観測されるスコアを求め、さらに、各モデルのスコアのうち、最も値の大きいもの（最高スコア）を求めて、ステップＳ４に進む。

【００３２】ステップＳ４では、認識処理部５は、最高スコアが所定の閾値εよりも大（所定の閾値以上）であるかどうかによって、センサ部１に入力された音声および画像のモデルが学習済みであるかどうかを判定する。ここで、モデルが学習済みかどうかの判定は、その他、例えば、音声認識において、入力音声が、辞書に登録されていない単語（未知語(OOV(Out Of Vocabulary))）かどうかを判定する手法等を適用して行うことが可能である。

【００３３】ステップＳ４において、最高スコアが所定の閾値εよりも大であると判定された場合、認識処理部５は、センサ部１に入力された音声および画像のモデルが学習済みであるとして、ステップＳ５に進む。ステップＳ５では、認識処理部５は、最高スコアのモデルに対応付けられた概念情報を、辞書データベース６の辞書から読み出し、認識結果として出力して、処理を終了する。

【００３４】一方、ステップＳ４において、最高スコアが所定の閾値εよりも大でないと判定された場合、認識処理部５は、いま認識の対象となっている音声および画像のモデルが学習済みでないとして、学習部７に対して、そのモデルの学習を要求して、ステップＳ６に進む。

【００３５】ステップＳ６では、学習部７は、メモリ４に記憶された合成特徴パラメータを用いて、センサ部１に入力された音声および画像のモデルの学習を行う。なお、モデルとしてＨＭＭを採用する場合には、モデルの学習には、例えば、Baum-Welchの再推定法などを用いることが可能である。

【００３６】学習部７は、学習によってモデルを得ると、そのモデルが表す音声および画像の概念情報の入力を要求するメッセージを、図示せぬディスプレイまたはスピーカより出力し、ユーザが、入力部８を操作することにより、概念情報を入力するのを待って、ステップＳ７に進む。

【００３７】ステップＳ７では、学習部７は、学習の結果得られたモデルと、入力部８が操作されることにより入力された概念情報と対応付けて、辞書データベース６の辞書に追加登録し、処理を終了する。

【００３８】なお、上述の場合には、最高スコアと閾値εとの大小関係に基づいて、認識処理部５から認識結果を出力し、あるいは、学習部７において学習を行うようにしたが、認識処理部５から認識結果を出力するか、あるいは、学習部７において学習を行うかは、その他、例えば、装置において、認識モードと学習モードとを切り替えるスイッチ等を設け、そのスイッチの操作に基づいて決定することが可能である。

【００３９】次に、図２のステップＳ１において同期処理部２が行う同期処理について説明する。

【００４０】例えば、いま、「ボールを蹴る」という音声と、ボールを蹴っている状態の画像とを、センサ部１から入力するとした場合、「ボールを蹴る」という音声が存在する区間（音声区間）と、ボールを蹴っている状態の画像（ボールを蹴るという行動が行われている画像）が存在する区間（画像区間）とは、図３に示すように、同期した状態にない。

【００４１】即ち、「ボールを蹴る」という音声の音声区間の始点（時刻）S_bと、ボールを蹴っている状態の画像の画像区間の始点M_bとは、一般に一致せず、「ボールを蹴る」という音声の音声区間の終点（時刻）S_eと、ボールを蹴っている状態の画像の画像区間の終点M_eも、一般に一致しない。従って、その音声区間の長さT_S（＝S_e－S_b）と、画像区間の長さT_V（＝M_e－M_b）も一致しない。

【００４２】また、上述のように、認識モードと学習モードとをスイッチによって切り替える場合に、学習モードにおいて学習を行おうとするときには、その学習対象の音声および画像は、例えば、図４に示すように、繰り返し入力されることがある。この場合も、図３における場合と同様に、繰り返し入力される音声の音声区間と、画像の画像区間とは、同期した状態にはならない。

【００４３】即ち、いま、ｉ番目に入力される音声と画像を、それぞれS_iとM_iと表すとともに、音声S_iの音声区間の始点、終点、長さを、それぞれＢ（S_i），Ｅ

（S_i），T（S_i）と表し、画像M_iの画像区間の始点、終点、長さを、それぞれB（M_i），E（M_i），T（M_i）と表すと、i番目の音声区間の始点B（S_i）と画像区間の始点B（M_i）は、一般に一致せず、終点E（S_i）とE（M_i）も、一般には一致しない。従って、音声区間の長さT（S_i）と、画像区間の長さT（M_i）も、一般には一致しない。

【００４４】このように、音声区間と画像区間の始点、終点、長さが一致しないと、音声と画像の特徴パラメータを合成した合成特徴パラメータにおいて、音声または画像の特徴パラメータのうちのいずれか一方が存在しない（いずれか一方だけ存在する）区間が生じ、この場合、ある同一の概念について、音声および画像の両方が入力されているのに、その概念の認識に、音声または画像のうちの一方だけしか用いられず、その結果、音声および画像の両方が用いられる場合に比較して、認識性能が劣化することとなる。

【００４５】そこで、図１の同期処理部２は、図５に示すように、センサ部１から供給される音声Ｓと画像Ｍを同期させる。

【００４６】即ち、同期処理部２は、例えば、図５（Ａ）に示すように、音声Ｓの音声区間の始点S_b'と、画像Ｍの画像区間の始点M_b'とを、ある時刻に一致させるとともに、音声区間の終点S_e'と、画像区間の終点M_e'とを、他の時刻に一致させる正規化処理を行う。なお、正規化が行われることにより、音声区間の長さと、画像区間の長さとは、一致することになる。

【００４７】より具体的には、同期処理部２は、例えば、音声の音声区間の始点S_b'、または画像Ｍの画像区間の始点M_b'のうちのいずれか一方を基準点とし、その基準点に、音声区間の始点S_b'と、画像区間の始点M_b'を一致させる。さらに、同期処理部２は、例えば、２００，４００、または８００ミリ秒等の所定の時間を基準時間長とし、音声区間と画像区間の長さが、いずれも基準時間長に一致するように、音声区間の終点S_e'と、画像区間の終点M_e'を変更する。従って、音声区間の終点S_e'と、画像区間の終点M_e'は、基準点から基準時間長だけ経過した点において一致することになる。

【００４８】さらに、同期処理部２は、正規化された音声区間の各音声フレームと、画像区間の各画像フレームとを、例えば、図５（Ｂ）に示すように、一対一に対応させる対応付け処理を行い、これにより、音声と画像とを同期させる。

【００４９】なお、正規化処理によれば、音声区間または画像区間の長さが変更することから、その音声区間または画像区間を構成する音声フレームまたは画像フレームの数を増減させる必要がある。また、対応付け処理においても、音声フレームと画像フレームの時間長が異なる場合には、音声フレームまたは画像フレームの数を増減させる必要があることがある。この音声フレームまたは画像フレームの数の増減は、例えば、補間や間引き等によって行うことが可能である。

【００５０】ここで、上述の場合には、音声フレームと画像フレームとを、一対一に対応付けるようにしたが、音声フレームと画像フレームとは、一対多または多対一に対応付けることも可能である。即ち、例えば、画像フレームの時間長が、音声フレームの時間長のＬ倍に一致する場合には、１フレームの画像フレームと、Ｌフレームの音声フレームとを対応付けるようにすることが可能である。

【００５１】次に、図６は、図１の同期処理部２の構成例を示している。

【００５２】センサ部１（図１）が出力する画像と音声は、区間検出部１１とメモリ１２に供給される。

【００５３】区間検出部１１は、そこに供給される画像の画像区間と、音声の音声区間とを検出し、区間正規化部１３および同期化部１４に供給する。即ち、区間検出部１１は、例えば、各画像フレーム全体の動きベクトルを求め、その動きベクトルに基づき、ある程度の動きのある画像フレームが連続している区間を画像区間として検出する。また、区間検出部１１は、例えば、各音声フレームのパワーを求め、そのパワーに基づき、ある程度のパワーを有する音声フレームが連続している区間を音声区間として検出する。

【００５４】メモリ１２は、そこに供給される画像データと音声データを一時記憶する。

【００５５】区間正規化部１３は、区間検出部１１から供給される音声区間を構成する音声フレームの音声データと、同じく区間検出部１１から供給される画像区間を構成する画像フレームの画像データを、メモリ１２から読み出す。さらに、区間正規化部１３は、その音声区間の音声データと、画像区間の画像データについて、上述した正規化処理を行い、これにより、始点、終点、長さを一致させた（正規化された）音声区間と画像区間の音声データと画像データを得て、同期化部１４に供給する。

【００５６】同期化部１４は、区間正規化部１３からの音声データと画像データについて、上述の対応付け処理を行うことにより、正規化された音声区間と画像区間の音声フレームと画像フレームとを一対一に対応付け、特徴抽出部３（図１）に出力する。

【００５７】次に、図７のフローチャートを参照して、図６の同期処理部２が、図２のステップＳ１において行う同期処理について説明する。

【００５８】センサ部１（図１）が出力する音声データと画像データは、区間検出部１１とメモリ１２に供給され、メモリ１２は、その音声データと画像データを一時記憶する。

【００５９】区間検出部１１は、ステップＳ１１において、センサ部１からの画像データの画像区間と、音声デ

ータの音声区間とを検出し、区間正規化部１３および同期化部１４に供給して、ステップＳ１２に進む。

【００６０】ステップＳ１２では、区間正規化部１３は、区間検出部１１からの音声区間と画像区間それぞれを構成する音声フレームの音声データと、画像フレームの画像データを、メモリ１２から読み出し、正規化処理を施すことで、始点、終点、長さを一致させた音声区間と画像区間、即ち、正規化された音声区間と画像区間の音声データと画像データを得て、同期化部１４に供給する。

【００６１】同期化部１４は、ステップＳ１３において、区間正規化部１３からの、正規化された音声区間と画像区間の音声フレームと画像フレームとを一対一に対応付ける対応付け処理を行うことで、音声フレームと画像フレームとをフレーム同期化し、特徴抽出部３（図１）に出力して、同期処理を終了する。

【００６２】以上のように、センサ部１から入力される、同一概念を表す音声データと画像データとを同期させるようにしたので、その音声データと画像データから得られる特徴パラメータを合成した合成特徴パラメータにおいて、音声または画像の特徴パラメータのうちのいずれか一方が存在しない区間が生じなくなり、そのような合成特徴パラメータを用いて認識が行われる結果、認識性能を向上させることができる。

【００６３】次に、上述した一連の処理は、ハードウェアにより行うこともできるし、ソフトウェアにより行うこともできる。一連の処理をソフトウェアによって行う場合には、そのソフトウェアを構成するプログラムが、汎用のコンピュータ等にインストールされる。

【００６４】そこで、図８は、上述した一連の処理を実行するプログラムがインストールされるコンピュータの一実施の形態の構成例を示している。

【００６５】プログラムは、コンピュータに内蔵されている記録媒体としてのハードディスク１０５やＲＯＭ１０３に予め記録しておくことができる。

【００６６】あるいはまた、プログラムは、フロッピー（登録商標）ディスク、CD-ROM(Compact Disc Read Only Memory)，MO(Magneto optical)ディスク，DVD(Digital Versatile Disc)、磁気ディスク、半導体メモリなどのリムーバブル記録媒体１１１に、一時的あるいは永続的に格納（記録）しておくことができる。このようなリムーバブル記録媒体１１１は、いわゆるパッケージソフトウエアとして提供することができる。

【００６７】なお、プログラムは、上述したようなリムーバブル記録媒体１１１からコンピュータにインストールする他、ダウンロードサイトから、ディジタル衛星放送用の人工衛星を介して、コンピュータに無線で転送したり、LAN(Local Area Network)、インターネットといったネットワークを介して、コンピュータに有線で転送し、コンピュータでは、そのようにして転送されてくるプログラムを、通信部１０８で受信し、内蔵するハードディスク１０５にインストールすることができる。

【００６８】コンピュータは、CPU(Central Processing Unit)１０２を内蔵している。CPU１０２には、バス１０１を介して、入出力インタフェース１１０が接続されており、CPU１０２は、入出力インタフェース１１０を介して、ユーザによって、キーボードや、マウス、マイク等で構成される入力部１０７が操作等されることにより指令が入力されると、それにしたがって、ROM(Read Only Memory)１０３に格納されているプログラムを実行する。あるいは、また、CPU１０２は、ハードディスク１０５に格納されているプログラム、衛星若しくはネットワークから転送され、通信部１０８で受信されてハードディスク１０５にインストールされたプログラム、またはドライブ１０９に装着されたリムーバブル記録媒体１１１から読み出されてハードディスク１０５にインストールされたプログラムを、RAM(Random Access Memory)１０４にロードして実行する。これにより、CPU１０２は、上述したフローチャートにしたがった処理、あるいは上述したブロック図の構成により行われる処理を行う。そして、CPU１０２は、その処理結果を、必要に応じて、例えば、入出力インタフェース１１０を介して、LCD(Liquid CrySta1 Display)やスピーカ等で構成される出力部１０６から出力、あるいは、通信部１０８から送信、さらには、ハードディスク１０５に記録等させる。

【００６９】ここで、本明細書において、コンピュータに各種の処理を行わせるためのプログラムを記述する処理ステップは、必ずしもフローチャートとして記載された順序に沿って時系列に処理する必要はなく、並列的あるいは個別に実行される処理（例えば、並列処理あるいはオブジェクトによる処理）も含むものである。

【００７０】また、プログラムは、１のコンピュータにより処理されるものであっても良いし、複数のコンピュータによって分散処理されるものであっても良い。さらに、プログラムは、遠方のコンピュータに転送されて実行されるものであっても良い。

【００７１】なお、図１の認識装置は、例えば、エンタテイメント用のロボットや、音声対話システム等に適用可能である。図１の認識装置をロボットに適用した場合には、例えば、ユーザの発話（例えば、「ボールが転がっている」など）と、対応する画像（例えば、ボールが転がっている様子が撮影された画像）から、周囲で生じている事象を、より正確に表す概念を得て、ロボットに的確な行動をとらせること等が可能となる。また、図１の認識装置を対話システムに適用した場合には、例えば、ユーザの発話とジェスチャから、ユーザの意図を、より正確に表す概念を得て、的確な応答を返すようにすること等が可能となる。

【００７２】ここで、本実施の形態においては、音声と

画像の両方から、それらが表す概念を認識するようにしたが、音声または画像のいずれか一方だけから、それが表す概念を認識するようにすることも可能である。

【００７３】また、センサ部１には、マイクとビデオカメラだけでなく、例えば、圧力センサ等の、画像および音声以外の外部からの刺激を感知するセンサを設け、そのセンサで得られる情報も、合成特徴パラメータに含めて、認識および学習を行うようにすることが可能である。

【００７４】

【発明の効果】本発明の認識装置および認識方法、並びに第１の記録媒体によれば、入力された画像と音声を同期させ、その同期された画像と音声それぞれから、特徴量を抽出して、その画像と音声の特徴量を合成した合成特徴量を得る。そして、その合成特徴量と、辞書におけるモデルとを用いてマッチングを行うことにより、入力された画像と音声が表す概念を認識する。従って、認識性能を向上させることが可能となる。

【００７５】本発明の学習装置および学習方法、並びに第２の記録媒体によれば、入力された画像と音声を同期させ、その同期された画像と音声それぞれから、特徴量を抽出して、その画像と音声の特徴量を合成した合成特徴量を得る。そして、その合成特徴量に基づいて学習を行うことによりモデルを生成し、同一概念を表す画像および音声に対応するモデルと、その画像および音声の概念を表す概念情報とを対応付けた辞書を生成する。従って、その辞書を用いた概念の認識を行う場合において、その認識性能を向上させることが可能となる。

【図面の簡単な説明】

【図１】本発明を適用した認識装置の一実施の形態の構成例を示すブロック図である。

【図２】認識装置の処理を説明するフローチャートである。

【図３】音声と画像が同期していない様子を示す図である。

【図４】音声と画像が同期していない様子を示す図である。

【図５】音声と画像が同期している様子を示す図である。

【図６】同期処理部２の構成例を示すブロック図である。

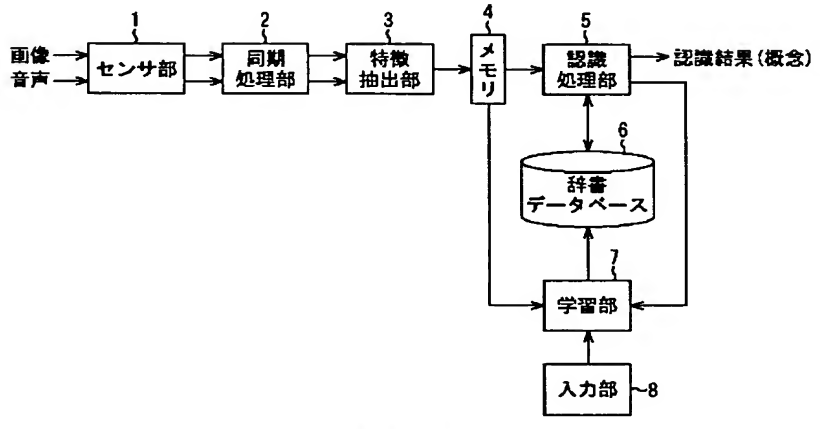
【図７】同期処理部２による同期処理を説明するフローチャートである。

【図８】本発明を適用したコンピュータの一実施の形態の構成例を示すブロック図である。

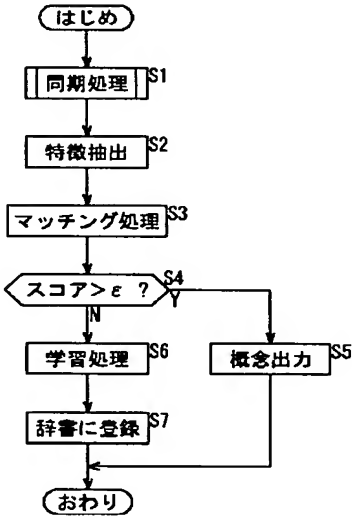
【符号の説明】

１　センサ部，　２　同期処理部，　３　特徴抽出部，　４　メモリ，　５認識処理部，　６　辞書データベース，　７　学習部，　８　入力部，　１１区間検出部，　１２　メモリ，　１３　区間正規化部，　１４　同期化部，　１０１　バス，　１０２　CPU，　１０３　ROM，　１０４　RAM，　１０５　ハードディスク，　１０６　出力部，　１０７　入力部，　１０８　通信部，　１０９　ドライブ，　１１０　入出力インタフェース，　１１１　リムーバブル記録媒体

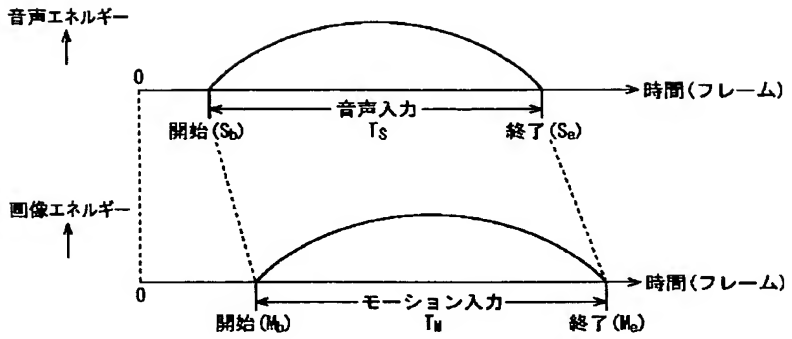
【図１】



【図２】



【図３】

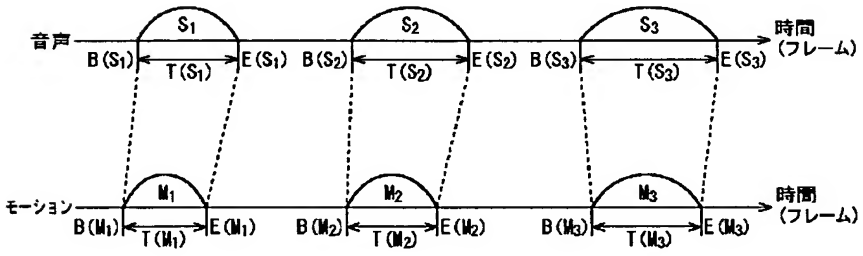


$S_b \approx M_b$

$S_e \approx M_e$

$T_S \approx T_M$ $(T_S = S_e - S_b,\ T_M = M_e - M_b)$

【図４】

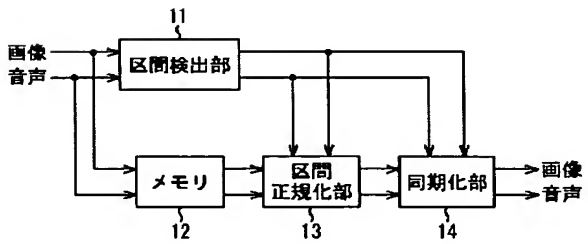


$B(S_i) \approx B(M_i)$

$E(S_i) \approx E(M_i)$

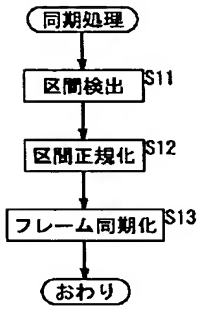
$T(S_i) \approx T(M_i)$

【図６】

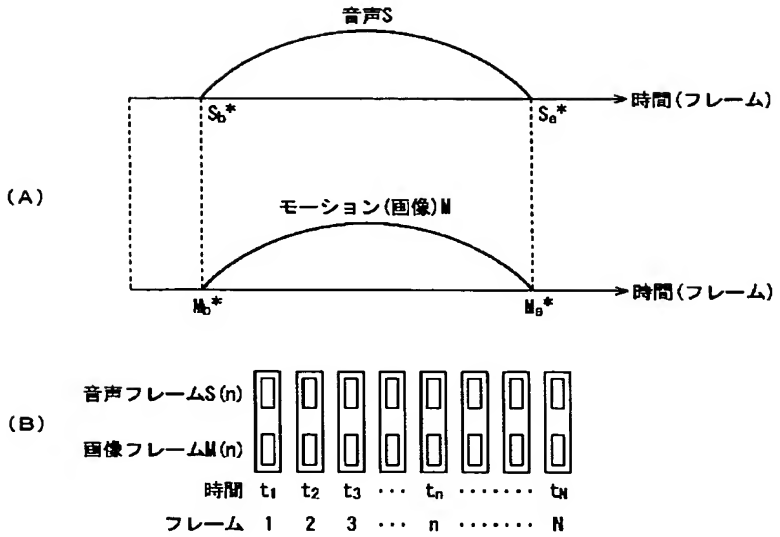


同期処理部 2

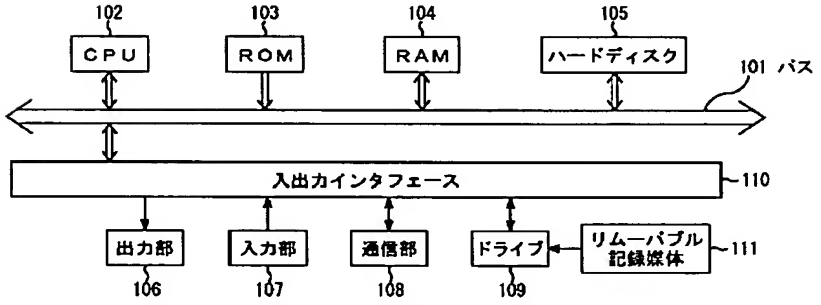
【図７】



【図５】


(A)
音声S
時間(フレーム)
S_b*
S_e*
モーション(画像)M
時間(フレーム)
M_b*
M_e*
(B)
音声フレームS(n)
画像フレームM(n)
時間 t_1 t_2 t_3 ・・・ t_n ・・・・・・・ t_N
フレーム 1 2 3 ・・・ n ・・・・・・・ N

【図８】


102
CPU
103
ROM
104
RAM
105
ハードディスク
101 バス
入出力インタフェース
110
出力部
106
入力部
107
通信部
108
ドライブ
109
リムーバブル記録媒体
111

コンピュータ